Semantic web-mining and deep vision for lifelong object discovery

Young, Jay; Kunze, Lars; Basile, Valerio ; Cabrio, Elena ; Hawes, Nicholas; Caputo, Barbara

DOI:
10.1109/ICRA.2017.7989323

Semantic Web-Mining and Deep Vision for Lifelong Object Discovery

Jay Young[1], Lars Kunze[1], Valerio Basile[2], Elena Cabrio[2], Nick Hawes[1] and Barbara Caputo [3]

Abstract— **Autonomous robots that are to assist humans in their daily lives must recognize and understand the meaning of objects in their environment. However, the open nature of the world means robots must be able to learn and extend their knowledge about previously unknown objects on-line. In this work we investigate the problem of unknown object hypotheses generation, and employ a semantic web-mining framework along with deep-learning-based object detectors. This allows us to make use of both visual and semantic features in combined hypotheses generation. Experiments on data from mobile robots in real world application deployments show that this combination improves performance over the use of either method in isolation.**

I. Introduction

Mobile service robots deployed in human environments such as offices, homes, industrial workplaces and similar locations must be equipped with ways of representing, reasoning and learning about the objects in their environment. Equipping a robot *a priori* with a (necessarily closed) database of object knowledge is difficult, because the system designer must predict which subset of all possible objects is required, and then build these models (a time-consuming task). If a new object appears in the environment, or an unmodelled object becomes important to a task, the robot will be unable to perceive, or reason about it. A solution to this problem is to give robots the ability to extend their own knowledge-bases on-line using information about new objects they encounter, and the capability to build models based on their own situated experiences. But it is not enough for a robot to merely learn a perceptual model of a newly observed cluster of 3D points or 2D pixels. Some form of *semantic* information is desirable too – for instance, how does it *relate* to other objects in the environment, where it might be *found*, what it is *used for* and where *should it go*. We refer to the linking of semantic knowledge to a previously unknown visual object as *hypothesis generation*.

While perceptual information about objects can be learned directly by a robot platform from its own situated observations [1], the question of how these observations are linked to semantic information is less clear. We expect that structured and semi-structured Web sources such as Wikipedia, DBPedia and WordNet [2] can be used to answer some of these questions.

[1]Intelligent Robotics Lab, University of Birmingham, United Kingdom {j.young,kunzel,n.a.hawes}@cs.bham.ac.uk

[2]Institut national de recherche en informatique et en automatique, WIMMICS, France {valerio.basile,elena.cabrio}@inria.fr

[3]University de Roma, Sapienza, Italy caputo@dis.uniroma1.it

A data source of particular interest to us is ImageNet, which is a large, ever-evolving database of categorised images organised using the WordNet lexical ontology. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [3] has in recent years produced machine learning tools trained on ImageNet for object detection and image classification. Of particular interest to us are *deep learning* based approaches using Convolutional Neural Networks, trained on potentially thousands of categories [4].

We expect such large-scale object detectors to be valuable in our hypotheses generation task as they provide a potential bridge between a robot's situated experience of objects and their associated semantic information. However, such an approach raises the question of how well such predictors perform when queried with the challenging image data endemic to mobile robot platforms, as opposed to the cleaner, and higher-resolution, data they are typically trained and evaluated on. Also, while this does not entirely address the problem of which objects to model in advance – using a CNN trained on ImageNet is still using a pre-trained detector, just one with a very large training set – the potential benefits are large, and would still allow us to extend a robot's knowledge base far beyond what it can be manually equipped with in advance of a deployment. Further, ImageNet is always growing and improving, so a robot's knowledge base could grow as new objects are added to it.

In this paper we investigate how semantic web-mining and deep vision can be combined to generate semantic label hypotheses for objects detected in real environments. These label hypotheses are linked to *structured, semantic knowledge bases* such as DBPedia and WordNet, allowing us to link perceptual experience with higher-level knowledge. This paper makes the following contributions:

- A novel approach for predicting the semantic identity of unknown, everyday objects based on web-mining using distributional semantics and Deep Vision.
- A surface-based approach to object learning on mobile robot platforms.
- An evaluation of our technique on real-world robot perception data from two long-term deployments.
- Provision of the software tools used to produce this work as open source software.

II. Related Work

To obtain information about unknown objects from the Web, a robot can use perceptual and/or knowledge-based queries. In this paper we use both types of queries. Knowledge-based queries can be seen as complementary to image-based queries which search databases of labelled

images for similarity, e.g. [5], or use web services such as Google Goggles to extract text, logo, and texture information [6], [7].

Although the online learning of new *visual* object models is currently a niche area in robotics, some approaches do exist [1], [8]. These approaches are capable of segmenting previously unknown objects in a scene and building models to support their future re-recognition. However, this work focuses purely on apperance models, and does not address how the learnt objects are described semantically. In a more supervised setting, many approaches have used humans to train mobile robots about new objects in their environment [9] and robots have also used Web knowledge sources to improve their performance in closed worlds, e.g. the use of object-room co-occurrence data for room categorisation in [10].

Our predictions for unknown objects rely on determining the semantic relatedness of terms. This is an important topic in several areas, including data mining, information retrieval and web recommendation. [11] applies ontology-based similarity measures in the robotics domain. Background knowledge about all the objects the robot could encounter, is stored in an extended version of the KNOWROB ontology [12]. Then, WUP similarity [13] is applied to calculate relatedness of the concept types by considering the depth of the concepts and the depth of their lowest common super-concept in the ontology. [14] presents an approach for computing the semantic relatedness of terms using ontological information extracted from DBpedia for a given domain, using the results for music recommendations. We compute the semantic relatedness between objects in mined text by leveraging the vectorial representation of the DBpedia concepts provided by the NASARI resource [15]. This method links back to earlier distributional semantics work (e.g. Latent Semantic Analysis [16]) with the difference that here concepts are represented as vectors, rather than words.

III. TASK DESCRIPTION AND DATA COLLECTION

In our work we consider a mobile service robot – in our case a MetraLabs Scitos G5, equipped with an ASUS Xtion RGB-D camera – tasked with observing everyday scenes in an unprepared human environment. By unprepared we mean that these are organically occurring scenes which we have not altered the scenes in anyway. We do not use the term "unstructured" because often these scenes have a natural structure that we wish to exploit.

Our robot is provided a map of the deployment environment, and each day it generates tasks to observe pre-selected cabinet tops, kitchen counters and other surfaces we determined to be potentially interesting for an object-learning robot. In the deployment environment we selected 30 surfaces of potential interest. Given a surface to observe, the robot takes multiple views from various angles, currently limited to 3. Our views are chosen using our ViPER library[1]

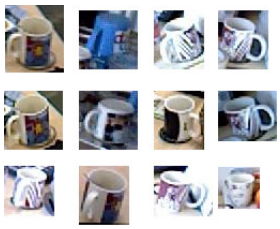

Fig. 1. A selection of mugs encountered by the robot.

which generates candidate views of surfaces in a stochastic way, and selects a limited set that aims to maximise coverage of the area. On each view, we segment the scene using a standard RANSAC-based plane-pop-out segmentation algorithm to remove table-like planes from the 3D point cloud, and cluster the results using the DBSCAN clustering algorithm, while applying size and luminance filters to filter out noisy segments. The thresholds for these filters were learned by gathering feedback from the robot's co-inhabitants regarding the quality of objects discovered by the robot after several weeks of deployment.

After taking multiple views of a surface, we achieve coherence between any objects that have been segmented by simply measuring the overlap between clusters, and linking clusters that have the most overlap, in what can be seen as a kind of 3D blob tracking. This allows us to link multiple views of a single object into an aligned co-ordinate space using the algorithms provided by the meta-room toolkit [17].

IV. SEMANTIC OBJECT KNOWLEDGE REPRESENTATION

Object models learned by the robot are stored in a long-term memory using our separately developed SOMa software[2], a database intended for representing and annotating objects in a robot's world model, augmented with a queryable spatio-temporal store.

In this paper we assume the robot is tasked with observing objects in unprepared, human environments. Whilst this is not a service robot task in itself, it is a precursor to many other task-driven capabilities such as object search, manipulation, human-robot interaction etc. Similar to prior work (e.g. [11]) we assume that the robot already has a semantic map of its environment which provides it with at least 3D models of supporting surfaces (desks, worktops, shelves etc.), plus the semantic category of the area in which the surface is located (office, kitchen, meeting room etc.). Surfaces and locations are linked to DBpedia, typically as entities under the categories `Furniture` and `Room` respectively.

A. Static and Dynamic Object Context

We split the spatial-semantic context representation of an object into static and dynamic parts. For the static part, we annotated local landmarks on our map of the environment, restricted to large non-moving objects such as coffee machines, photocopiers, sinks, fax machines, whiteboards, and other objects that are typically stationary on a day-to-day

[1]https://github.com/kunzel/viper

[2]http://github.com/strands-project/soma

basis. Also part of the static representation is a label such as `Kitchen` or `Office` that broadly describes the usage of a particular area. For the dynamic part, we included any object detected from a pre-trained set taken from the deployment environment and learned manually using the V4R toolkit[3], which also provides our recognition software.

V. Semantic Web-Mining

In previous work we developed a Semantic Web-Mining component for robot systems [18]. This component provides access to object- and scene-relevant knowledge extracted from Web sources, and is accessed using JSON-based HTTP requests. The structure of a request to the system describes the objects that were observed with an unknown object, the spatial relations held between each object, calculated exhaustively pairwise, as well as the room and surface labels describing where in the environment the observations were made. Upon receiving a query, the service computes the *semantic relatedness* between each object included in the co-occurrence structure and every object in a large set of candidate objects from which possible concepts are drawn from. This semantic relatedness is computed by leveraging the vectorial representation of the DBpedia concepts provided by the NASARI resource [15]. The NASARI resource represents BabelNet concepts [19] as a vector in a high-dimensional geometric space. The vector components are computed with the *word2vec* [20] tool, based on the co-occurrence of mentions of each concept, in this case using Wikipedia as source corpus. Using the *distributional hypothesis*, vectors that represent related entities end up close in the vector space, allowing us to measure relatedness by computing the inverse of the cosine distance between two vectors. For instance, the NASARI vectors for `Pointing_device` and `Mouse_(computing)` have relatedness 0.98 (on a continuous scale from 0 to 1), while `Mousepad` and `Teabox` are 0.26 related. The system computes the aggregate of the relatedness of a candidate object to each of the scene objects contained in the query. Formally, given n observed objects in the query $q_1, ..., q_n$, and m candidate objects in the universe under consideration $o_1, ..., o_m \in O$, each o_i is given a score that indicates its likelihood of being the unknown object by aggregating its relatedness across all observed objects. The aggregation function we use to give the likelihood of an object o_i is given by:

$$likelihood(o_i) = \prod_{j=1}^{n} relatedness(o_i, q_j)$$

We also make use of *Qualitative Spatial Relations* (QSRs) to represent information about objects [21]. QSRs discretise continuous spatial measurements, particularly relational information such as the distance and orientation between points, yielding symbolic representations of ranges of possible continuous values. For more details see [18].In this work, we make use of a qualitative distance measure, often called a Ring calculus. When observing an object, we categorise its distance relationship with any other objects in a scene with the following set of symbols: $near_0, near_1, near_2$, where $near_0$ is the closest. This is accomplished by placing sets of thresholds on the distance function between objects, and in this way, spatial proximity of observed objects is taken into account in the context representation used to find semantically related object.

[3]https://github.com/strands-project/v4r

VI. Deep Vision

Deep learning in general is being explored more and more by the robotics community, being used for such tasks as grasp detection [22] and visual perception tasks [23]. In this work, we make use of Deep Convolutional Neural Networks trained on data from the ImageNet project of crowd-sourced annotated images. Specifically we make use of the CNN architecture of Krizhevsky et. al [4], which is implemented in the Caffe toolkit [24]. This is attractive to us, as such predictors are trained on an extensive amount of data, but such models are relatively small and computationally cheap to query. In our system we use this as a predictor, and pass in cropped images of objects the robot discovers autonomously. In return, the CNN provides a ranked list of object label hypotheses as WordNet classes.

VII. Combining Semantic Web-Mining and Deep Vision

We seek to combine the predictive power of our Semantic Web-mining approach and existing Deep Vision techniques into a single system capable of generating object label hypotheses. We also wish this to be possible in a *multi-view* way, where multiple observations of a single object are made from different vantage points.

Algorithm 1 describes our approach in pseudo code. The algorithm takes the following arguments as its input (Line 1): a cropped image of the target object I; two lists of labels which determine the dynamic (C_D) and the static (C_S) context in the scene; and two parameters n and t which control the filtering process (here we used n=7 and t=.8). The algorithm returns a label for the target object l^*. First, we initialize the set of candidate labels (Line 3). We then predict a set of labels and their corresponding confidences from the cropped image I using a trained CNN (Line 4). After selecting the n best labels from the CNN (Line 5), we iterate over these labels, and relate them to all context labels by computing the WUP score (Line 8). If the WUP score is larger than the predefined threshold t (Line 9), we add the CNN label to the list of candidates (Line 10). Eventually we select the label with the highest confidence from all candidates (Line 14) and return it (Line 15).

As a worked example, in one occasion the robot observes an object for which the ground truth is Microwave. Our CNNs provide us a ranked list of potential labels, the top ranked of which being `Safe` (as in a safety deposit box), the second being `Crate`, with confidences 0.47 and 0.34 respectively, the third being `Microwave` with a confidence of 0.27, and the final two predictions being `Fire screen` (0.12) and `Screen` (0.013). Our context system predicts

Algorithm 1: Object Label Prediction based on Semantic Web-Mining and Deep Vision

Function predictObjectLabel (I, C_D, C_S, n, t)
Input : Cropped image I; Dynamic context C_D; Static context C_S; Number of CNN and context candidates n; Semantic relatedness threshold t
Output: Label l^*
begin
  $L_{Candidates} \leftarrow \emptyset$
  $L_{CNN} \leftarrow predictLabelsFromCNN(I)$
  $L'_{CNN} \leftarrow selectNBestLabels(L_{CNN}, n)$
  for $l \in L'_{CNN}$ **do**
    for $c \in (C_D \cap C_S)$ **do**
      $wup \leftarrow computeWUP(l, c)$
      if $wup > t$ **then**
        $L_{Candidates} \leftarrow L_{Candidates} \cap\ l$
      end
    end
  end
  $l^* \leftarrow \arg\max_l L_{Candidates}$
  return l^*
end

various kitchen appliances and objects, given that the object is seen in a room labelled Kitchen, and there are objects like a kettle, a mug and a fridge nearby. We find that in the context predictions suggestions such as `Oven` and `Toaster` relate strongly to the `Microwave` prediction of the CNN (WUP 0.92 in both cases), where the `Crate` and `Safe` predictions do not relate strongly to any of the context predictions (WUP 0.52 and 0.50 respectively), nor do the `Fire screen` or `Screen` predictions (again 0.52 and 0.50 respectively, indicating these entries are distant). As such, filtering means we drop all hypotheses from the CNN results except for Microwave, which we then put forward as the label for this view – if we had multiple possible candidates after filtering, we would defer to their original CNN ranking and pick the highest. In the case of there being no candidate that is above the relatedness threshold, we will always defer to the vision system, and we see this occurs in 23.40% of cases in dataset B and 37.9% of cases in dataset A. Upon taking multiple views however, we find that occasionally the label `Dishwasher` wins out instead, being suggested by the CNN with a high confidence and *also* being highly related to our context predictions. Overall however this error is lessened by the combination of votes from multiple views – more views vote for the `Microwave` label than the `Refrigerator` label overall – meaning that our prediction of the object's label is overall correct in this particular encounter.

VIII. Experiments

We evaluate our system on two datasets, both collected by the same robot platform in two separate, large workplace environments across two deployment episodes. Dataset A was collected using our surface-based object learning system described previously in this paper, and dataset B was collected using the meta-room paradigm of [17]. The core difference between these modalities is that our approach is based on *directed* views onto *specific* surfaces, whereas the meta-room approach is largely a brute-force approach, making $360°$scans of areas in the deployment environment. Both datasets were labelled by hand, and along with this paper we make our labelling tool available to the community as Open Source software [4]. Dataset B contains around 2000 views of individual objects. For each object instance, there are in general around 2-5 views, of the 18 objects in the experiment on average this gives us between 20-25 instances. Since the deployment environment was an active, working office, some objects – such as monitors – are overrepresented in the data, due to there being many in the environment. The least common objects were the Fire Extinguisher and the Microwave, as the robot did not as often visit the small kitchen in which they reside.

In our experiments, our measure of success is the WUP similarity [13] between the prediction of a system and the ground truth object label. To do so we map our DBPedia-based labels from the context prediction system to their equivalent WordNet classes, as this is the domain used by ImageNet. Certainly we could have used the single vectorial representation throughout – rather than just for the measurement of co-occurence in the web-mining component, as we do. We chose instead to use WordNet concept distance when calculating the accuracy of predictions against ground truth. We found this measure to be less noisy than a DBPedia distance in constraining the meaning of objects, and operates on more abstract entity types than DBPedia (Bluetooth Keyboard, Cordless Keyboard, Ergonomic Keyboard in DBPedia VS. Just *Keyboard* in WordNet). Objects in our system maintain links to *both* representations, allowing us to make use of information found in one resource but not the other, and vice-versa, in the future. WUP similarity is one standard measure of calculating the semantic relatedness of word senses in the lexical ontology of WordNet. A WUP score of 1.0 means two concepts are identical. For instance, in WordNet the concepts *dog* and *cat* have a WUP score of 0.86, a *computer keyboard* and a *mouse* have WUP score of 0.80, a *laptop computer* and a *cow* have a WUP score of 0.30. We regard any WUP score above 0.70 as indicating a good categorical relation.

Our use of the WUP score, as opposed to a binary true/false accuracy measure, is because we are interested in predictions that are *strongly categorically related* to the true identity of unknown objects. We do not view the system described in this paper in isolation, and consider it as an important component in an overall, integrated approach to unknown object identification for mobile robots. In our next steps, a reasonable list of hypotheses will allow us to boost our ability to employ more specific, and potentially expensive, methods, for selecting from a set of hypotheses towards

[4] http://github.com/jayyoung/lwann

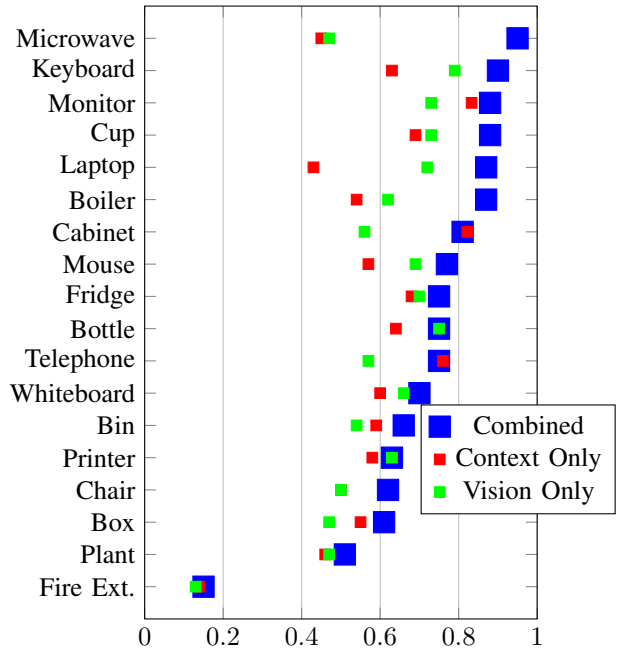

Fig. 2. WUP Values between Ground Truth and Predicted Object Label for DataSet B, gathered using the meta-room approach.

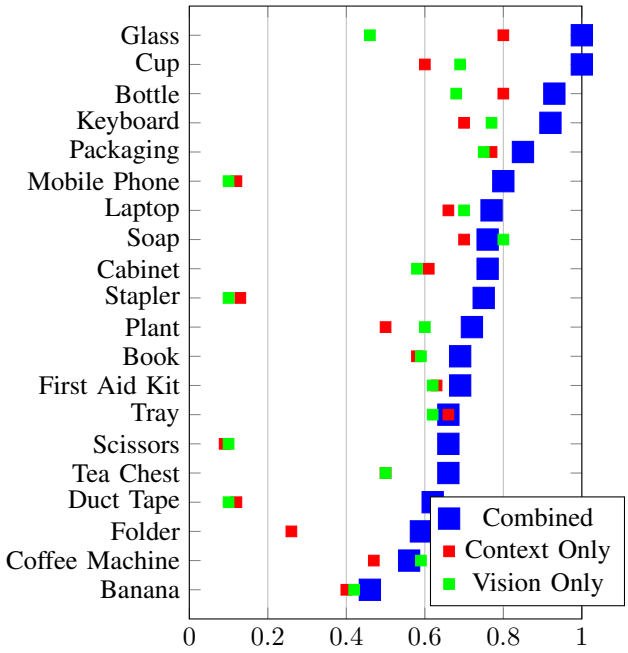

Fig. 3. WUP Values between Ground Truth and Predicted Object Label for DataSet A – gathered using our surface-based approach with view planning.

extending its knowledge-base. For instance, the robot may present the objects it has discovered in the environment to its human co-inhabitants and ask for their help in refining its hypotheses. But in order to do so intelligently and efficiently, a set of *initial* hypotheses is crucial.

For experiments with dataset B we employed a leave-one-out approach, where the context for a given object is provided as the other objects in the scene. This provides parity with previous experiments [18]. In experiments with dataset A, we use the static context and results of pre-trained object recognisers described previously.

IX. Results

The results of our experiments are shown in Figures 2 and 3.

In general our results show that our system is able to effectively constrain label hypotheses, and in several cases is either entirely accurate (as with the Drinking Glass and Cup seen in dataset A), or comes extremely close (As with the Microwave, Keyboard and Monitor in dataset B). However we observe that accuracy does go down the more noisy the views – in the case of Dataset A, the robot only encountered 3 mugs during its time in the environment, taking a total of 11 views. On the other hand, there are over 160 views across 30 instances of various types of cups in Dataset B.

The results highlight multiple interesting problems with our approach. The first of which being that, since we had no control over the classes the CNN was trained with, we could not guarantee that the specific objects the robot observed in the environment would be detectable. In dataset B we see that the *Fire Extinguisher* is the object that all systems have the most trouble with – the reason why the CNN performs so poorly is that it does not appear to have been trained with any images similar to the the kind of Fire Extinguisher the robot observed in the environment. The reason why the context-based system fails is because the Fire Extinguisher was observed in a kitchen, amongst other typical kitchen utensils and objects, and a strong relationship was unable to be found between the objects that could be observed and a Fire Extinguisher. This highlights a flaw in our context-based approach in that it struggles to identify objects that are *surprising* or out of place – ideally in such a case we would hope for the vision system to rectify this since, as discussed previously, if no label passes our relatedness test we fall back to the predictions of the CNN entirely. But here, both systems fail. The same is true for the Banana in Dataset A – it is observed in a workspace environment amongst various pieces of computer hardware, and so the context system constraints its predictions to those kinds of objects. The inability to cope properly with objects that are out-of-context with their environment is a significant limitation of our system. What is particularly frustrating is that the CNN *does* typically recognise bananas when they are observed, and reports a confidence of around 0.70, which is excellent and can be put down to the distinct shape and texture of the fruit.

X. Discussion

We evaluated our system on real-world unprepared scenes, which were necessarily noisy, diverse, and featured dynamics such as occlusion and varied lighting conditions. We know from our own experimentation that the CNNs we used perform very well on clear, high-resolution images of objects similar to those observed by our robot. But their perfor-

mance on robot data is significantly worse, and we believe this highlights the limitations of several of our approaches, and underpins the problem of domain adaptation. Certainly typical arguments can be made that the performance of such vision systems can be improved by using higher-resolution sensors, and we do agree that easy gains can be made in this way. But this only serves to paper over a very interesting problem – a *human* is able to identify objects from the kind of noisy, low-resolution data our robot has collected, and so should a robot platform. We believe that *robot vision* must be treated as its own distinct area of research, where the problems of perception and action must be addressed in an integrated way, taking into account the specific dynamics and problems associated with mobile platforms. The best way to work towards that goal is to develop and evaluate robot-centric algorithms and techniques, and evaluate them in situated, real-world scenarios, rather than scenarios where the designer has influenced the experimental set-up and may unconsciously introduce bias or reduce noise and dynamics.

XI. Conclusion

We presented a system that allows a mobile robot to generate label hypotheses for unknown objects it encounters in its environment. We used a semantic web-mining system that allows us to generate candidate labels for an object given it's spatial-semantic context, and coupled this with a deep vision system trained on ImageNet. Using the context predictions as a preference heuristic to select a subset of the predictions made by the deep vision algorithm, we tested this system on data gathered by a robot operating in two real-world workplace environments, and using two different modalities of data collection. Our results showed that we are able to effectively constrain the set of possible labels for a given object using our approach, though large variance in performance is seen depending on the uniqueness of the object and the availability of trained classifiers. This work highlights the crucial need for integrated approaches to robot perception, and for those techniques to be evaluated on *situated*, real-world data.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No 600623, STRANDS, and under the ALOOF project (CHIST-ERA program).

References

[1] T. Faeulhammer, R. Ambrus, C. Burbridge, M. Zillich, J. Folkesson, N. Hawes, P. Jensfelt, and M. Vincze, "Autonomous learning of object models on a mobile robot," *IEEE RAL*, vol. PP, no. 99, pp. 1–1, 2016.

[2] A. Kilgarriff and C. Fellbaum, "Wordnet: An electronic lexical database," 2000.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[5] J. Philbin, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR 2008*, June 2008, pp. 1–8.

[6] M. Beetz, F. Bálint-Benczédi, N. Blodow, D. Nyga, T. Wiedemeyer, and Z.-C. Marton, "Robosherlock: Unstructured information processing for robot perception," in *ICRA*, 2015.

[7] D. Nyga, F. Balint-Benczedi, and M. Beetz, "PR2 Looking at Things: Ensemble Learning for Unstructured Information Processing with Markov Logic Networks," in *IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, May 31-June 7 2014.

[8] R. Finman, T. Whelan, M. Kaess, and J. J. Leonard, "Toward lifelong object segmentation from change detection in dense rgb-d maps," in *ECMR*. IEEE, 2013.

[9] G. Gemignani, R. Capobianco, E. Bastianelli, D. Bloisi, L. Iocchi, and D. Nardi, "Living with robots: Interactive environmental knowledge acquisition," *Robotics and Autonomous Systems*, 2016.

[10] M. Hanheide, C. Gretton, R. Dearden, N. Hawes, J. L. Wyatt, A. Pronobis, A. Aydemir, M. Göbelbecker, and H. Zender, "Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour," in *IJCAI'11*, Barcelona, Spain, July 2011.

[11] M. Schuster, D. Jain, M. Tenorth, and M. Beetz, "Learning organizational principles in human environments," in *ICRA*, May 2012, pp. 3867–3874.

[12] M. Tenorth and M. Beetz, "KnowRob – A Knowledge Processing Infrastructure for Cognition-enabled Robots," *Int. Journal of Robotics Research*, vol. 32, no. 5, pp. 566 – 590, April 2013.

[13] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *ACL*, ser. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, pp. 133–138.

[14] J. P. Leal, V. Rodrigues, and R. Queirós, "Computing Semantic Relatedness using DBPedia," in *1st Symposium on Languages, Applications and Technologies*, ser. OpenAccess Series in Informatics (OASIcs), A. Simões, R. Queirós, and D. da Cruz, Eds., vol. 21. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012, pp. 133–147.

[15] J. Camacho-Collados, M. T. Pilehvar, and R. Navigli, "Nasari: a novel approach to a semantically-aware representation of items." in *HLT-NAACL*, R. Mihalcea, J. Y. Chai, and A. Sarkar, Eds. The Association for Computational Linguistics, 2015, pp. 567–577.

[16] T. K. Landauer and S. T. Dutnais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *PSYCHOLOGICAL REVIEW*, vol. 104, no. 2, pp. 211–240, 1997.

[17] R. Ambruş, N. Bore, J. Folkesson, and P. Jensfelt, "Meta-rooms: Building and maintaining long term spatial models in a dynamic world," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 1854–1861.

[18] J. Young, V. Basile, L. Kunze, E. Cabrio, and N. Hawes, "Towards lifelong object learning by integrating situated robot perception and semantic web mining," in *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2016.

[19] R. Navigli and S. P. Ponzetto, "Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, no. 0, pp. 217 – 250, 2012.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[21] L. Frommberger and D. Wolter, "Structural knowledge transfer by spatial abstraction for reinforcement learning agents," *Adaptive Behavior*, vol. 18, no. 6, pp. 507–525, Dec. 2010.

[22] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[23] A. Giusti, J. Guzzi, D. C. Cireşan, F.-L. He, J. P. Rodríguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. Di Caro, *et al.*, "A machine learning approach to visual perception of forest trails for mobile robots," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 661–667, 2016.

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.